Visualisierung von lexikalischem Wandel im Deutschen auf Basis der Google-Books Ngram Daten

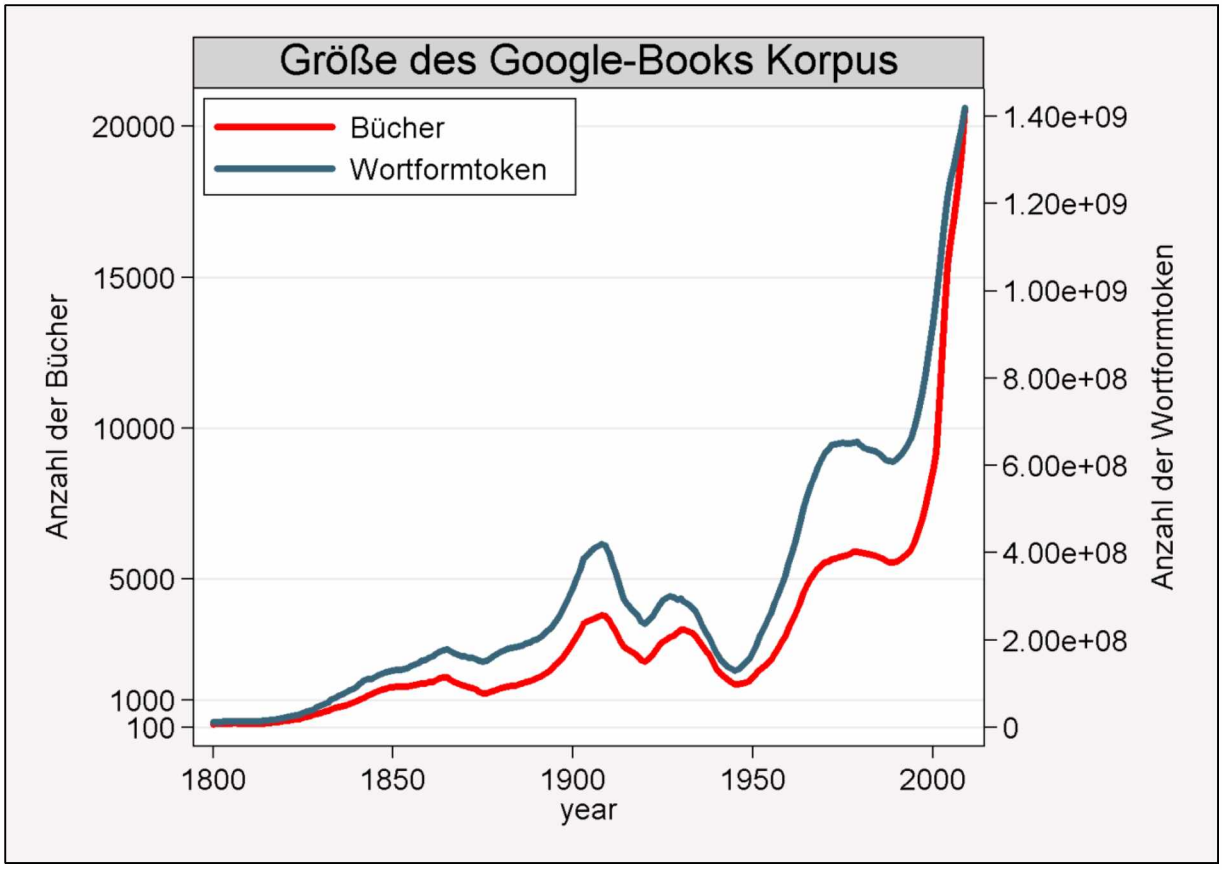
Arbeitspapier

Zusammenfassung

In diesem Arbeitspapier wird gezeigt, wie mit Hilfe der Google-Books Ngram Daten (Michel u. a., 2010a, 2010b) lexikalischer Sprachwandel visualisiert werden kann.

Daten

Zunächst wurden mit Hilfe der unter http://storage.googleapis.com/books/ngrams/books/datasetsv2.html vorhandenen Datensätze alle deutschen Unigramm-Datensätze in der Version 20120701 heruntergeladen. Diese Datensätze wurden in einem nächsten Schritt zu einem Datensatz zusammengeführt, welcher für jede vorhandene Wortform für jedes Jahr die jeweilige absolute Korpushäufigkeit und die Anzahl der Nennungen in verschiedenen Büchern auflistet. Zusätzlich wurden die vorhandenen Wortklasseninformationen verwendet (Lin u. a., 2012). So findet sich zum Beispiel das Substantiv „Korpus“ im Jahr 1964 81 Mal in sechs verschiedenen Büchern. Grafik 1 visualisiert die Korpusgröße im zeitlichen Verlauf[1], sowohl in Bezug auf die im jeweiligen Jahr vorhandenen Bücher also auch in Bezug auf die im jeweiligen Jahr enthaltenen Wortformtoken.



Grafik 1: Größe des Google-Books Korpus im zeitlichen Verlauf

[1] Die Daten beruhen auf folgender Datei: http://storage.googleapis.com/books/ngrams/books/googlebooks-ger-all-totalcounts-20120701.txt.

Anmerkung: Um zufällige Schwankungen aus den Daten herauszurechnen und um so langfristige zeitlichen Trends visualisieren zu können, wurden alle Daten für diese und alle anderen in diesem Papier präsentierten Grafiken geglättet, indem für jeden Wert ein lokaler Mittelwert berechnet wurde, welcher neben dem eigentlichen Wert alle Werte in einem Fenster von ±5 Jahren beinhaltet. Der Wert für das Jahr 1900 ist somit der Mittelwert aus den Jahren 1895, 1896, 1897, 1898, 1899, 1900, 1901, 1902, 1903 und 1904 (Becketti, 2013, S. 100–137).

Datenanalyse

Für die Analyse des lexikalischen Wandels wurden zunächst Wortformtypen der Klassen Numerale und Satzzeichen, sowie Abkürzungen und Fremdwörter (Wortklasse „andere“) aus den Daten getilgt. Um den stark variierenden Korpusgrundlagen Rechnung zu tragen (vgl. Grafik 1), wurde für jeden Wortformtyp die relative Häufigkeit pro eine Millionen Wortformtoken berechnet und Wortformtypen, die seltener als ein Mal pro eine Millionen Wortformtoken in den Daten auftreten, aus der Analyse ausgeschlossen.

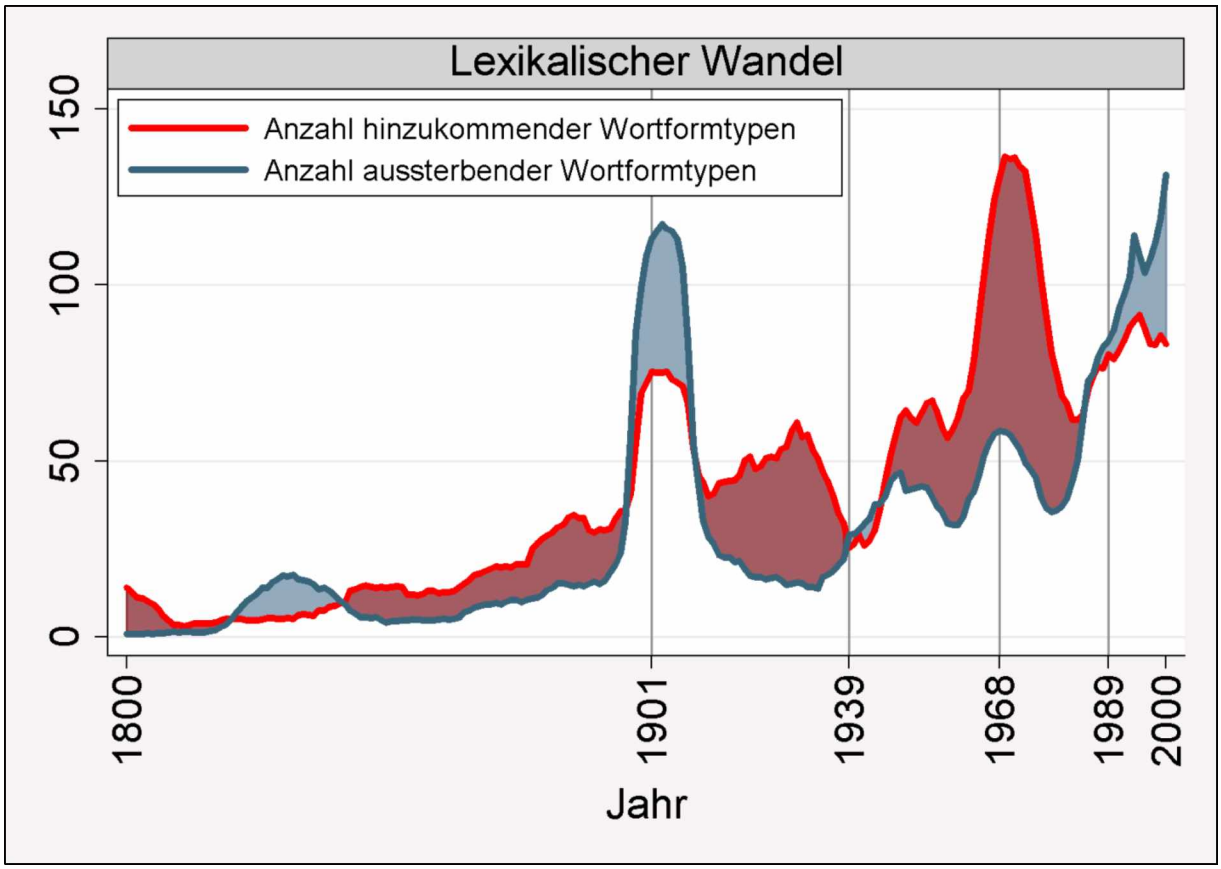
Basierend auf diesem Datensatz wurde dann im Zeitraum von 1751 bis 2009 für jedes Jahr berechnet, (a) wie viele Wortformtypen im dem jeweiligen Jahr neu hinzugekommen sind und (b) wie viele Wortformtypen in dem jeweiligen Jahr verschwunden sind.

Anmerkung Um Zufallsbefunde aus den Daten herauszurechnen, wurden die beiden Kenngrößen wie folgt operationalisiert:

(a) Als Geburt werden nur Wortformtypen gezählt, die in dem jeweiligen Jahr zum ersten Mal belegt sind und die zusätzlich in *mindestens acht der zehn folgenden Jahre* ebenfalls belegt sind.

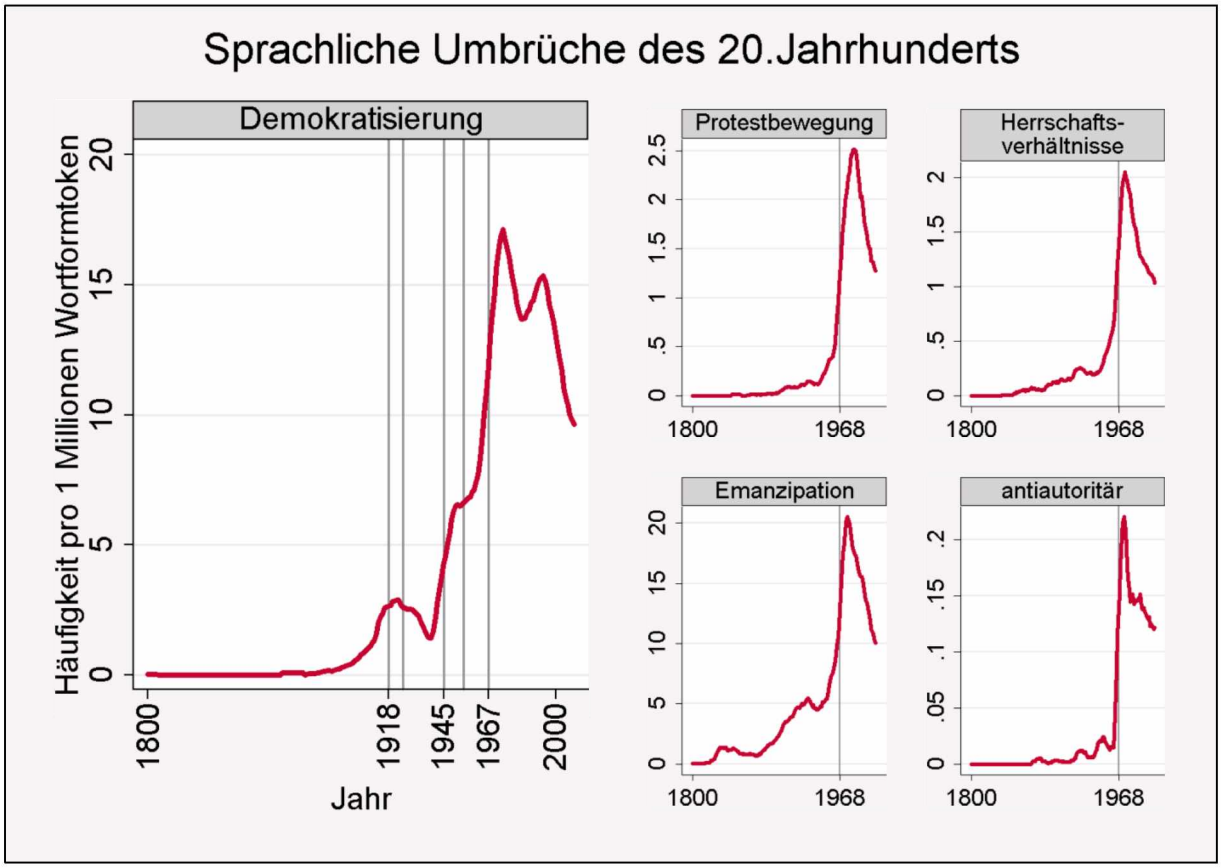
(b) Als Aussterben werden nur Wortformtypen gezählt, die in dem jeweiligen Jahr zum letzten Mal belegt sind und die zusätzlich in *mindestens acht der zehn vorherigen Jahre* belegt waren.

Grafik 2 visualisiert die Ergebnisse.



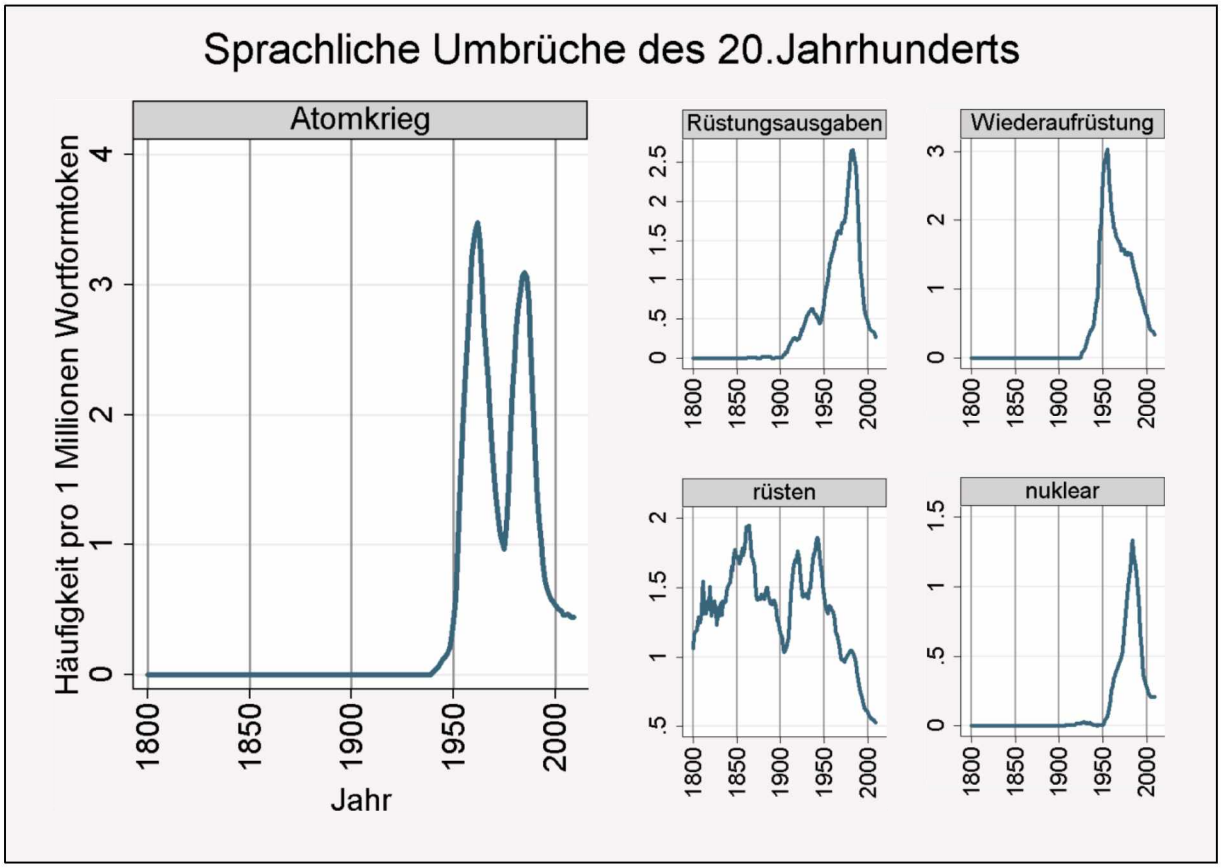
Grafik 2: Lexikalischer Wandel basierend auf dem Google-Books Korpus

Um die Ergebnisse noch anschaulicher gestalten zu können, wurde in einem weiteren Schritt einige Wortformtypen extrahiert, welche in dem Diskurswörterbuch „Protestdiskus 1967/68“ näher beschrieben wurden (vgl. Kämper, 2012, sowie http://www.owid.de/wb/disk68/start.html). Grafik 3 zeigt, wie sich die relative Häufigkeit der betreffenden Wortformtypen als Indikator für dessen gesellschaftliche Relevanz gewandelt hat.



Grafik 3: Sprachliche Umbrüche des 20.Jahrhunderts am Beispiel von ausgewählten Beispielen

Grafik 4 zeigt abschließend, wie mit Hilfe des hier skizzierten Verfahrens auf zusätzliche gesellschaftliche Prozesse geschlossen werden kann, für welche sich lexikografische Projekte unter Umständen lohnen könnten.



Grafik 4: Sprachliche Umbrüche des 20.Jahrhunderts am Beispiel von ausgewählten Beispielen

Literatur

Becketti, S. (2013). *Introduction to time series using Stata* (1st ed.). College Station, Tex: Stata Press.

Lin, Y., Michel, J.-B., Aiden, L. E., Orwant, J., Brockmann, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (S. 169–174). Jeju, Republic of Korea.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Verses, A., Gray, M. K., The Google Books Team, … Aiden, L. E. (2010a). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(14), 176–182. doi:10.1126/science.1199644

Michel, J.-B., Shen, Y. K., Aiden, A. P., Verses, A., Gray, M. K., The Google Books Team, … Aiden, L. E. (2010b). Quantitative Analysis of Culture Using Millions of Digitized Books (Supporting Online Material). *Science*, *331*(14). doi:10.1126/science.1199644